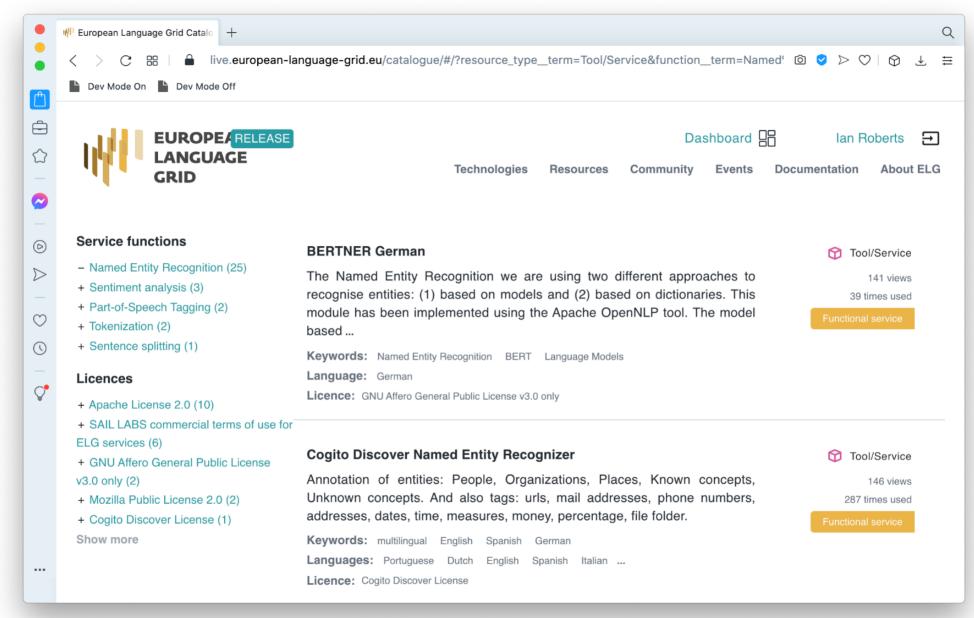


European Language Grid – META-FORUM 2020 Kalina Bontcheva – European Language Grid: Service and Resources

- Current State of Play
- Future Plans: Releases 2 and 3



## LT Services: Current state of play

- Release 1 in April 2020 finalized APIs for major classes of services (ASR, IE, MT, TTS)
- Concentrated on a subset of EU languages Czech, English, French, German, Greek, Latvian, Spanish (native languages of the ELG consortium)
  - 9 ASR services the seven priority languages plus Estonian and Lithuanian
  - ~150 distinct services for IE & Text Analytics 52 English, 28 German, 21 French, 14 Greek
    - Fewer actual catalogue entries as one endpoint often serves multiple functions
  - 24 MT the priority languages to/from English, plus a few others supported "for free" by the same systems
  - 2 TTS Latvian and Lithuanian

#### **ELG R1: Services**

#### IE and Text Analytics

Etiquetas de fila	ch	English	French	German	Greek	Latvian	Spanish	Total general
<b>■CUNI</b>	6	; <b>6</b>	5 5	5	5	5	5	37
Dependency Parsing	1	. 1	. 1	. 1	. 1	. 1	1	. 7
Lemmatisation	1	. 1	. 1	. 1	. 1	. 1	1	. 7
Morphological analyser	1	. 1	. 1	. 1	. 1	. 1	1	. 7
Named Entity Recognition	1	. 1						2
Part of Speech tagging	1	. 1	. 1	. 1	. 1	. 1	1	. 7
Tokenization	1	. 1	. 1	. 1	. 1	. 1	1	. 7
<b>■ DFKI</b>	1	. 8	3 2	. 6	1	. 1	2	21
Categorization		1						1
Language identification	1	. 1	. 1	. 1	. 1	. 1	1	. 7
Morphological analyser		1	. 1	. 1			1	. 4
Named Entity Recognition		2	!	2				4
Sentence splitting		1		1				2
Summarization		1						1
Tokenization		1		1				2
<b>■ Expert System</b>	1	. 7	' 6	5 5	1	. 1	6	27
Categorization		1					1	. 2
Language identification	1	. 1	. 1	. 1	. 1	. 1	1	. 7
Lemmatisation		1	. 1	. 1			1	. 4
Named Entity Recognition		1	. 1	. 1			1	. 4
Part of Speech tagging		1	. 1	. 1			1	. 4
Sentiment Analysis		1	. 1					2
Summarization		1	. 1	. 1			1	. 4
■ILSP		1			4			5
Information Extraction					2			2

**■ CUNI** Czech-English English-Czech **English-French** French-English DFKI English-French English-German **English-Spanish ■ILSP** English-Greek Greek-English ■Tilde English-Bulgarian English-Latvian English-Polish Latvian-English Polish-English **UEDIN** Czech-English English-Czech English-German German-English

**ASR ■ SAIL LABS** English French German Greek Spanish ■Tilde Latvian **■ UEDIN** Czech

TTS ■Tilde Latvian Lithuanian

tiquetas de fila	▼ Czech	E	nglish F	rench G	erman G	ireek	Latvian	Spanish	Total genera
CUNI		6	6	5	5	5	5	5 5	3
Dependency Parsing		1	1	1	1	1	1	. 1	
Lemmatisation		1	1	1	1	1	1	. 1	
Morphological analyser		1	1	1	1	1	1	. 1	
Named Entity Recogniti	on	1	1						
Part of Speech tagging		1	1	1	1	1	1	. 1	
Tokenization		1	1	1	1	1	1	. 1	
DFKI		1	8	2	6	1	1	. 2	. 2
Categorization			1						
Language identification		1	1	1	1	1	1	. 1	
Morphological analyser			1	1	1			1	
Named Entity Recogniti	on		2		2				
Sentence splitting			1		1				
Summarization			1						
Tokenization			1		1				
Expert System		1	7	6	5	1	1	. 6	;
Categorization			1					1	
Language identification		1	1	1	1	1	1	. 1	
Lemmatisation			1	1	1			1	
Named Entity Recogniti	on		1	1	1			1	
Part of Speech tagging			1	1	1			1	
Sentiment Analysis			1	1					
Summarization			1	1	1			1	
ILSP			1			4			
Information Extraction						2			
Named Entity Recogniti	on		1			1			
Sentiment Analysis						1			
SAIL LABS		2	3	3	3	2		3	
Language identification		1	1	1	1	1		1	
Named Entity Recogniti	on	1	1	1	1	1		1	
Sentiment Analysis			1	1	1			1	
USFD		1	27	5	9	1	1	. 4	
Categorization			4						
Language identification			1	1	1			1	
Morphological analyser			1						
Named Entity Recogniti	on		7	2	4				
<b>NER Disambiguation</b>			4	1	1			1	
Number annotation			1						
Opinion Mining			2						
Part of Speech tagging		1	3	1	1	1	1	. 1	
Sentence splitting			1		1				
Summarization			1					1	
			2		4				
Tokenization			2		1				

# LT Services: Approaching Release 2

- Release 2 (February 2021) will add support for other EU and related languages, at least:
  - 8 additional ASR services
  - 200-250 additional IE & Text Analysis services
  - 23 additional MT services
  - 9 TTS services
- We also expect the first services and datasets from Pilot Projects to come on stream at or shortly after Release 2
  - More details later in the conference ...

# LT Services: Looking further ahead

- Release 3 in early 2022 will introduce services for an even wider range of non-EU languages
- Current projection is for at least
  - 15 further ASR
  - 160 further IE/Text Analysis
  - 9 further MT
- Further service types such as image OCR, terminology extraction from corpora, etc.

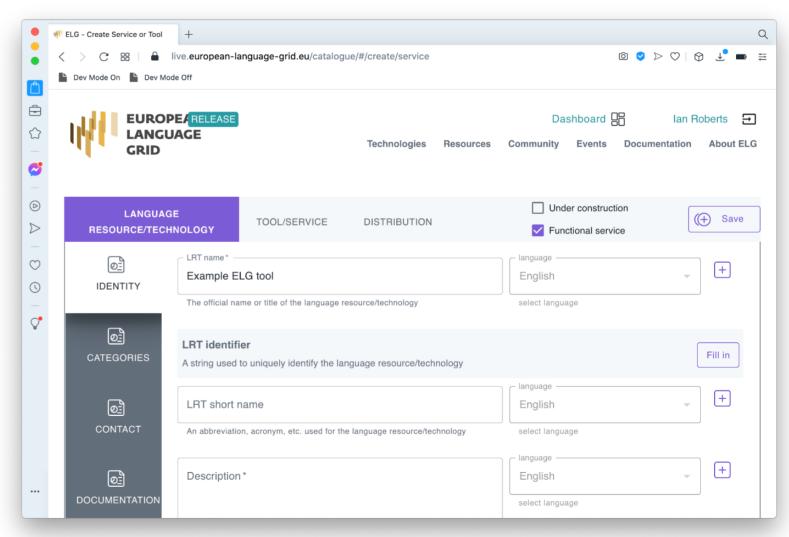


## Adding your own services to ELG

- Current tools have taken anything from a few hours to a few days to integrate
  - Some are easier than others
- Hope to get this down across the board to minutes in the future
- We have helper libraries that deal with much of the complexity, e.g.
  - Spring Boot Starter for Java you provide one implementation class, the rest is boilerplate
  - Python helper in development

## Adding your own services to ELG

- Metadata graphical editor just launched
  - Or XML upload to register services in bulk





#### Language Resources: Current state

- ELG Release 1 had approx. 280 freely available language resources integrated
- Since then:
  - Ingestion completed: ELRA, ELRC-SHARE, ELRA-SHARE-LRs (2014to2018) and LINDAT/CLARIAH-CZ
  - Other identified repositories in progress: Zenodo and Quantum Stat
- Different procedures depending on repository:
  - Metadata converters
  - Mapping and harvesting
  - Manual enriching of metadata
- Some resources are hosted within ELG, others just metadata referring back to source repository

# LRs available in ELG

	Corpora	Lexical and Conceptual Resources	Models and Computational Grammars	Total
ELRA	635	545	-	1180
ELRC-SHARE	844	43		887
META-SHARE	52	12	7	71
LINDAT/CLARIAH-CZ	243	66	-	309
ELRA-SHARE-LRs	46	25	-	71
Zenodo	36	37	-	73
Total	1857	727	7	2591



European Language Grid

## Next steps to populate the ELG catalogue

#### • META-SHARE:

- In progress managing nodes from consortium partners are being ingested into ELG
- Over 200 repositories have been identified:
  - · Zenodo:
    - 73 datasets ingested into ELG so far
    - About 600 dataset records are under analysis
  - Quantum Stat:
    - About 480 dataset records are under analysis
  - Other repositories are selected following LR gaps
- LRs from pilot projects

